IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-11, Nov 2018 E-ISSN: 2347-2693

Author Identification on Imbalanced Class Dataset of Indian Literature in Marathi

Sunil D. Kale[1*], Rajesh S. Prasad[2]

[1] Computer Engineering Department, Smt. Kashibai Navale College of Engineering, Pune, Maharashtra, India
[2] Computer Engineering Department, Sinhgad Institute of Technology and Science, Pune, Maharashtra, India

[*]Corresponding Author: kalesunild@gmail.com, Tel.: +91-9921932670

Available online at: www.ijcseonline.org

Accepted: 18/Nov/2018, Published: 30/Nov/2018

Abstract— Author Identification is one of the application of text mining and is the task of investigating author of the anonymous text document. Application of author Identification includes in digital forensic, plagiarism detection, copyright issues, etc. The numerous amount of work is already done on English language perhaps Author identification of Indian regional languages is limited. This research paper presents Author identification on Indian regional Marathi Language. In this paper proposing a technique for identifying probabilistic authors via linguistic stylometry i.e. the statistical analysis of variations in literary style between one author or genre with another. In total 11 features are extracted with 8 lexical and syntactic features and 3 word N-gram features. Experimentation is performed with 8 features and machine learning algorithms, i.e. k-nearest neighbor, Naïve Bayes and Support Vector Machine. Moreover, result based on word n-gram i.e. unigram, bigram and trigram are also presented. Experimentation result shows better result with word N-gram method.

***Keywords*—**Author Identification**,** Text Mining, Machine Learning, Marathi Language, Stylometry

I. INTRODUCTION

The process of identifying the author from a group of candidate author according to writing samples (sentences, paragraphs or short articles) is author identification [1]. Author Identification has great importance because of its wide applications. It can be used to find the original author of widely reprinted articles, as well as plagiarized paragraphs [1]. The authorship of a bit of writing is the identity of the individual that wrote it. There are numerous assignment are carrying out with authorship, including, authorship identification, verification, and profiling. In this paper , we focus on predicting the unknown authors of Marathi articles for stopping the unauthorized get admission to or usage of articles. India is a center of great heritage, it is basically union of different cultures which has a great impression all over the world. It is the land of great linguistic variety. There are several languages spoken in the country, but officially recognized are 22 languages [2]. Languages pronounced is an in India are from different language groups, the major ones being the Indo-Aryan languages spoken by 78.05% of Indians and the Dravidian languages spoken by 19.64% of Indians. Languages spoken by the remaining 2.31% of the population belong to the Austroasiatic, Sino-Tibetan, Tai-Kadai, and a few other minor language families and isolates. India has the world's second highest number of languages, after Papua New Guinea [3].Author Identification is done by identifying author's writing style. The writing style of the author is nothing but the text used by author for writing document which also termed as author's stylometry. We can find the writing style of author by identifying textual characteristics that he/she used while writing a document. Process of Identifying Author, Step I: Provide training data as text written by different authors. Step II: Identify feature set. Step III: Based on the features selected in Step II classify authors by applying some statistical computations and decide the similarity score of author. Step IV: Input test data i. e. anonymous text file that should belong to one of the authors used in training set, but not same text data as which was used in the training data. Step V: Apply Step III on test data it gives a similarity score of test text file. Step VI: Compare similarity scores of training text files and test text file which is calculated based on feature set count with some statistical computations. If a match is found, then one is the probable author of input anonymous text document.

Rest of the paper is structured as follows. In section 2, research done on Author Identification is provided. In section 3, proposed technique is shown. In section 4, details of text corpus used and feature extraction is provided. In section 5, experimentation and results are discussed and finally in section 6 conclusion and future scope is given.

II. Related Work

The application of stylometry is done for a long time on English contents. The first approach was by Mendenhall, 1877 on Shakespeare's work, by proposing the word-length distribution as a feature [4]. In the computer era, approach towards stylometry was viewed from a statistical perspective. This approach was set by Frederick Mosteller and David Wallace in 1964, who used function-words & Bayesian analysis for the identification of the author of the disputed Federalist papers [5]. Author identification on different languages a detailed survey provided by [6]. Morever Author Identification methods with systematic review is given in [7]. Researchers [8] proposed a method for identifying the author of Arabic text articles. In this research work researchers proposed an intelligent classifier that is capable of predicting the author of a piece of writing given a predefined set of candidate authors and a number of samples for each author. In classification, they utilized the Functional Trees (FT) and sequential minimization optimization which is a variant of Support Vector Machines (SVM). The function tree algorithm received 82% of accuracy during the classification of Arabic text author. In some cases, the SVM performed well and achieved 100% accuracy. [9] proposed an author identification system for Turkish Texts. In this research article, the researchers collected Turkish text documents from URL www.hurriyet.com.tr. First, researchers developed the Turkish directory and rules for the language. These texts have different subjects such as magazine, medical and politics. During the classification of Turkish author, they presented the experimentation in three methods. In the first method used all 22 features which are calculated as equal weights and achieved a success rate of 67%. In the second method with modification to first one, the success rate has improved approximately to 84%. [10] Presented an author identification system for Albanian text articles. This approach is based on the statistical measures and methods applied to rewrite rules which appear in a syntactically annotated corpus. Experimentation is performed in three different methods. In the first method, 31 of 43 books, author was successfully identified. In the second method, 38 of 43 books, author was successfully identified. The third method, 40 of 43 books, author was successfully identified. First experimentation on author identification on Indian regional language Marathi is done by [11], with sequential minimal optimization using proposed features specific for text in Marathi Language. [12] Presented an author identification method on Telugu text language. In this approach, the phases used for extracting the pattern include preprocessing, feature extraction, feature selection, classification and then at the end the author selects. The researchers of this article telling that "Unigram" has outperformed on Telugu text classification compared to any other kind of classy. SVM has been quite effective in obtaining accurate results by considering variety of classifiers which include Decision tree, K-nearest neighbor, and Nave byes classifier. Suprabhat [13] proposed an author identification method for Bengali articles. Researchers [14][15] provided feature extraction work for Guajarati Language text also presented template matching algorithm.

III. Proposed Technique

Considering the previous work and research done in the field of author identification and realizing the need for a experimentation on Indian regional Marathi language, in this section we propose a model for language spoken mostly is the western part of India, which is Marathi. Marathi Indo-Aryan language and it is one of the official languages of India. It has a great literature written in Devanagari script. Many well-known authors have their great contribution to the Marathi Literature authors like P.L Deshpande, V.P Kale, Ram Ganesh Gadkari, etc. In the following model we focus on identifying the writing style of the author and then find out the probable author of the test data. Figure 1 shows the proposed model. The model consists of three main sections: Preprocessing, Feature extraction and Classifier. In the first half the data of different authors are taken as input, this data is called as training data.

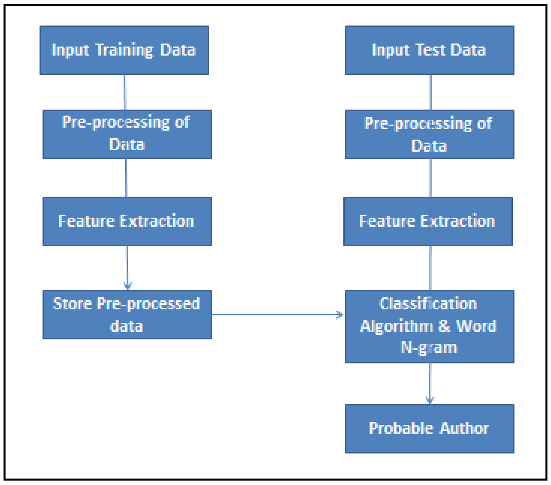


Figure 1. Proposed Model of the System.

Preprocessing: It is used for tokenization of words in a file. Here, different methods are used such as stemming, stop word removal. After preprocessing the data, it is stored in the database.

Feature Extraction: This is the core of the system, which help in distinguishing the writing style of different authors. Features are extracted after the preprocessing is stored in the database. Features such as word frequency, punctuations, n-gram method etc. are selected as shown in table 2. In the second half the same process of preprocessing and feature extraction is carried out on the test data. Then the classifier is applied on the extracted features.

Classifier: It is multiclass problem so used statistical classification, where classification is identifying to which group of category a new observation (test document) belongs, on the basis of the trained data in the database [16]. Classifiers used are K-Nearest Neighbor, Naïve Bayes and Support Vector Machine (SVM).

IV. TEXT CORPUS AND FEATURE EXTRACTION

A. *Text Corpus*

Text corpus collection is one of the most challenging tasks to accomplish for electronically constrained language Marathi. As per our knowledge, no standard text corpus is available for Author identification, so we have prepared text corpus from various Marathi literary websites. Text collection is done of 20 well known authors in Marathi. The collected text is imbalanced in size. The Text corpus consists of different texts which consist of Katha, Lekh, Natak, Lalit. The table 1 shows the name of authors number of documents by respective author and its size specifically to show the imbalance of data size. The Text corpus consists of a lot of Marathi literature available freely on the internet. The data collected is mostly in .txt form. Marathi Literature is very vast which consists of Natak, Kavita, Kadambari, lekh, etc. The Text corpus mentioned here is a mixture of categories of writing as a Katha, lekh, Natak and Lalit lekhan.

Table 1. Text corpus prepared for experimentation

Sr. No.	Name of author	No. of Documents	Total Size
1	Sameer	80	1.87 MB
2	Runmesh	45	520.1 KB
3	Atul Thakur	25	547.2 KB
4	Nandini Desai	22	1 MB
5	Mohana	21	521.8 KB
6	Akhildeep	20	848.56 KB
7	Mahendra Kulkarni	20	507.2
8	Vinit Dhanawade	16	2.25 MB
9	Priti Chatre	16	620.4 KB
10	P. L. Deshpande	16	917.4 KB
11	Shraddha Bhowad	15	767.6 KB
12	Ramesh Wagh	15	547.8 KB
13	Avinash Joshi	14	903.4 KB
14	Meghna Bhuskute	14	559.8 KB
15	Vidya Bhutkar	13	673.4 KB
16	Kavathi Chfa	12	578.2 KB
17	Koutuk Shirodhkar	12	666 KB
18	Aditi Joshi	10	720.8 KB
19	Aniket Pise	6	640.3 KB
20	Nimish Sonar	3	558.9 KB

B. *Feature Extraction*

Lexical features: A text is generally viewed as a sequence of tokens grouped together to form a sentence. Token can be a word, number or a punctuation mark. The first attempt made by Mendenhall, 1887, for author identification was based on sentence length count and word length count [17]. Advantages of such features are that they can be applied to each and every language irrespective of the structure of language. In the past, vast majority of author identification is based on lexical features. Some of these features used in the system are mentioned in table 2.

The field of stylometry is a development of literary stylistic's and can be defined as the statistical analysis of literary style [18]. It makes some basic assumption that any author has its own distinctive writing habits that eventually reflects in the features such as the author's core vocabulary usage, sentence complexity and phraseology that is used. Stylometry attempts to find features of an author's writing and to determine statistical methods to measure these features so that similarity between texts can be analyzed. Features are mainly the core of any writing, they give meaning to a structure of any textual data. They are useful for distinguishing writing style of different authors and genres. Most of the work is done mainly in English, but each language has its own richness and own way of structure. There are many possible features that can be used for identifying the author. The number of features and their type often varies in author identification, to determine the importance and influence of certain features. Table 2 mentioned below shows the different types of features used in our system.

Table 2. Features used for experimentation.

Sr. No.	Features	Description
1	Word Size	Count of Character
2	Sentence Length	Count of words in a line
3	Paragraph Lines	Count of no. of lines in a paragraph
4	Type Token Ration	Number of distinct words
5	Pace	Repetition of same words
6	Hapax Legomena	Words occurring only once
7	Hapax Dislegomena	Words occurring twice
8	Punctuation Marks	!, ?, ;, -, etc.
9	Word N-gram	Unigram, Bigram, Trigram.

Word Size - Word Length is calculated by considering each character of a word. Here, the word length of each file is calculated as,

$$Word\ Size = \frac{Total\ number\ of\ characters}{Total\ number\ of\ words}$$

Sentence Length - It is basically the number of words in a line. Here, the line length of each file is calculated as,

$$Scenetnce\ Length = \frac{Total\ number\ of\ words}{Total\ number\ of\ Lines}$$

Paragraph Lines - It is the number of lines in a paragraph, a paragraph is found by comparing the character with the newline (\n). Here, the paragraph length of each file is calculated as,

$$Paragraph\ Lines = \frac{Total\ number\ of\ lines}{Total\ number\ of\ paragraph}$$

Type Token Ratio (TTR) - This feature helps to find the distinct words in a text, the formula for its calculation is,

$$TTR = \frac{T}{N}$$

Where, T= Number of distinct tokens, the N = Total number of tokens of the text.
PACE - It is to find the number of times an author repeats the same token.

$$PACE = \frac{1}{TTR}$$

Hapax Legomena - Here, the words most rarely used are identified. The words that not occur more than once.

$$HLRT = \frac{T1}{N}$$

Where, HLR= Hapax Legomena Ratio, T1= number of once occurring tokens, N= Total number of tokens in a text.

Hapax Dislegomena - Here, the words occurring twice are considered and their respective frequency is calculated.

$$HDRT = \frac{T2}{N}$$

Where, HDR = Hapax Dislegomena Ratio, T2 = number of twice occurring tokens, N= Total number of tokens in a text.

V. Experimentation and Results

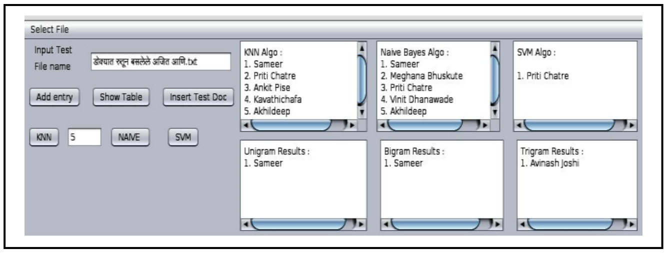
The data structure is actually a part of any experiment that is used for organizing and storing data in a computer, which can be further accessed and modified efficiently. The data structure used in the module is HashMap. HapMap is a map based collection class that is used for storing key and value pairs, denoted as Hashmap(key, value) [19]. The language used for programming is Java, thus Java HashMap class implements the map interface by using a hash table [20].

Experimentation carried out by partitioning text corpus into three sets as shown below:

Training set and Testing suite: - The training and testing text for each author are as follows:-

1. 70% of training and 30% Testing
2. 80% of training and 20% Testing
3. 90% of training and 10% Testing

The feature extraction is performed on the training and testing set. The features are mapped
In the algorithms used, which are K-Nearest Neighbor, Naïve Bayes and Support Vector Machine. Some most versatile features which directly help to identify the author of anonymous text are unigram, bigram and trigram. In Figure 2 the input unknown file naming डोक्यात रुतन बसलेले अजित आजि is given for testing purpose. The results we get to the file are correctly predicted by K-Nearest Neighbor, Naïve Bayes, Unigram and Bigram i.e. the author of the unknown corpus is Sameer. The following figure shows us the actual implementation and results of the mapping.



Input Test File name: डोक्यात रुतून बसलेले अजित आणि.txt

5

KNN Algo :
1. Sameer
2. Priti Chatre
3. Ankit Pise
4. Kavathichafa
5. Akhildeep

Naive Bayes Algo :
1. Sameer
2. Meghana Bhuskute
3. Priti Chatre
4. Vinit Dhanawade
5. Akhildeep

SVM Algo :
1. Priti Chatre

Unigram Results :
1. Sameer

Bigram Results :
1. Sameer

Trigram Results :
1. Avinash Joshi

Fig. 2. Experimentation and the corresponding results.
Performance Measures

Accuracy and relevance is one of the most important keys for any project model. The
Methods used to specify the relevance of the author predicted and the corresponding accuracy of the purposed model are specified below. Confusion Matrix is a specific table layout that helps us to find the performance measures. Each row of the matrix represents the predicted class while each column represents an actual class [21].
Precision: It is the number of correct results divided by the number of all returned results.

$$Precision = (TP + FP)$$

TP = the number of positive samples correctly predicted by the classification method.
FP = the number of negative samples incorrectly predicted as the positive class.

Recall: It is the number of correct results divided by the number of results that should have been returned.

$$Recall = TP(TP + FN)$$

FN = the number of positive samples incorrectly predicted by the classification method.

The figure 3 shows the accuracy of result achieved by K-NN, Naïve bayas and SVM. Figure 4 shows the accuracy of result achieved by word n-gram method.

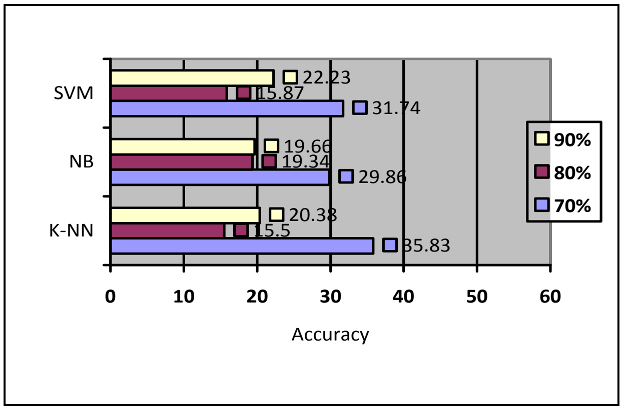


Fig. 3. Comparison of the accuracy of results based on training data size.

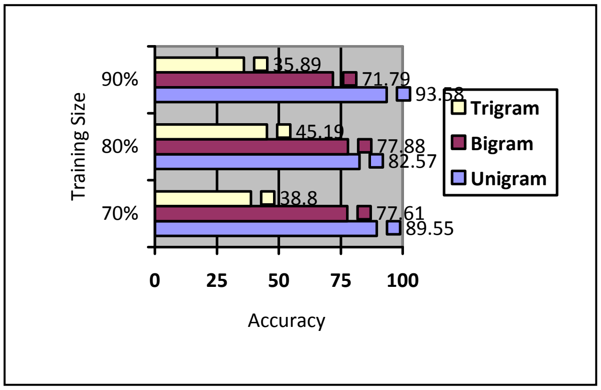


Figure 4:- Comparison of the accuracy of results based on word n-gram method with varying training data size.

F-Measure: A measure that combines precision and recall is the harmonic mean of precision and recall.

$$F - Measure = 2 * \frac{precision * recall}{precision + recall}$$

Accuracy: It is the statistical measure which tells how correctly the results are identified.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

N = the number of negative samples correctly predicted as the negative class.

VI. CONCLUSION AND FUTURE

The proposed technique for author identification on Marathi literature helps in finding the most relevant, probabilistic author of the anonymous text, more efficiently by word n-gram method specifically by the unigram, the accuracy achieved is 93.58% for training data of 90% and testing data of 10% as shown in the figure 4. Figure 3 shows that if the training data is of smaller size like 70%, then K- Nearest Neighbor gives the better accuracy as compared to the other two algorithms; similarly if the training data are of moderate size like 80%, then Naïve Bayes gives the better accuracy as compared to the other two algorithms; if the training data is of larger size like 90%, then Support Vector Machine gives the better accuracy as compared to the other two algorithms. The results of accuracy obtained through machine learning algorithms are not satisfactory.Future scope is to enhance the language specific feature set. Current world is of internet and everything is available over internet, which leads to criminal and malicious activity. So, the identity of available content is now a need. The available content is always in the form of text data. This hypothesis is surely helpful in reducing the increasing crimes, growing in and around our environment on a daily basis. Author identification is definitely a great help in digital forensics, copy-wright issues, and historical context recognition and computer security, making the law process quick and efficient.

The main conclusions of the study may be presented in a short Conclusion Section. In this section, the author(s) should also briefly discuss the limitations of the research and Future Scope for improvement.

REFERENCES

[1] C. Qian, T. He, and R. Zhang, "Deep Learning based Authorship Identification."

[2] Wikipedia contributors, "Languages with official status in India-Wikipedia," *Wikipedia, The Free Encyclopedia.*, 2018. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Languages_with_official_status_in_India&oldid=841744869. [Accessed: 21-May-2018].

[3] "Diversity of India – Geographical and Cultural contexts – Am an aspirant too," *Wikipedia, The Free Encyclopedia.* [Online]. Available: https://tklvch.wordpress.com/2015/01/07/diversity-of-india-geographical-and-cultural-contexts/. [Accessed: 27-Apr-2018].

[4] T. C. Mendenhall, "The characteristic curves of composition.," *Science*, vol. 9, no. 216, pp. 237–249, 1887.

[5] F. Mosteller and D. Wallace, "Inference and disputed authorship: The Federalist," 1964.

[6] K. S. Digamberrao and R. S. Prasad, "Author Identification on Literature in Different Languages: A Systematic Survey," in *2018*

International Conference On Advances in Communication and Computing Technology (ICACCT), 2018, pp. 174–181.

[7] S. D. Kale and R. S. Prasad, "A Systematic Review on Author Identification Methods," *Int. J. Rough Sets Data Anal.*, vol. 4, no. 2, pp. 81–91, Apr. 2017.

[8] A. F. Otoom, E. E. Abdullah, S. Jaafer, A. Hamdallh, and D. Amer, "Towards author identification of Arabic text articles," in *2014 5th International Conference on Information and Communication Systems (ICICS)*, 2014, pp. 1–4.

[9] B. Diri and M. Fatih Amasyali, "Automatic Author Detection for Turkish Texts."

[10] H. Paci, E. Kajo, E. Trandafili, I. Tafa, and D. Salillari, "Author identification in Albanian language," in *Proceedings - 2011 International Conference on Network-Based Information Systems, NBiS 2011*, 2011, pp. 425–430.

[11] S. D. Kale and R. S. Prasad, "Author Identification using Sequential Minimal Optimization with rule-based Decision Tree on Indian Literature in Marathi," *Procedia Comput. Sci.*, vol. 132, pp. 1086–1101, Jan. 2018.

[12] S. N. Prasad, V. B. Narsimha, P. V. Reddy, and A. V. Babu, "Influence of Lexical, Syntactic and Structural Features and their Combination on Authorship Attribution for Telugu Text," *Procedia Comput. Sci.*, vol. 48, no. C, pp. 58–64, 2015.

[13] S. Das and P. Mitra, "Author Identification in Bengali Literary Works," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6744 LNCS, springer, 2011, pp. 220–226.

[14] J. R. Prasad, U. V. Kulkarni, and R. S. Prasad, "Template Matching Algorithm for Gujrati Character Recognition," in *2009 Second International Conference on Emerging Trends in Engineering & Technology*, 2009, pp. 263–268.

[15] J. R. Prasad, U. V. Kulkarni, and R. S. Prasad, "Offline Handwritten Character Recognition of Gujrati script using Pattern Matching," in *2009 3rd International Conference on Anti-counterfeiting, Security, and Identification in Communication*, 2009, pp. 611–615.

[16] F. Wikipedia, "Statistical classification Frequentist procedures."

[17] E. Stamatatos, "A survey of modern authorship attribution methods," *J. Am. Soc. Inf. Sci. Technol.*, vol. 60, no. 3, pp. 538–556, 2009.

[18] M. W. Corney, "Analysing E-mail Text Authorship for Forensic Purposes by," 2003.

[19] Chaitanya Singh, "HashMap in Java with Example." [Online]. Available: https://beginnersbook.com/2013/12/hashmap-in-java-with-example/. [Accessed: 29-Oct-2018].

[20] "HashMap in Java - javatpoint." [Online]. Available: https://www.javatpoint.com/java-hashmap. [Accessed: 29-Oct-2018].

[21] E. Table, R. External, C. Cat, and D. Rabbit, "Confusion matrix," pp. 1–4, 2018.

Authors Profile

Sunil D. Kale has received his graduate degree from Government College of Engineering, Aurangabad, Maharashtra, India and master's degree M. Tech. (CSE) from Visvesvaraya Technological University, Karnataka, India. He is a Research Scholar in Computer Engineering Department of Smt. Kashibai Navale College of Engineering, Vadgaon (Bk) and working as Assistant Professor in Computer Engineering Department of Pune Institute of Compter Technoloy, Pune, India. His area of interest is text analytics and pattern recognition. He has published 17 papers in national and international journals and conferences. He is a life member of Indian Society of Technical Education.

Rajesh S. Prasad has received Masters (M.E. Computer Engg.) degree from College of Engineering, Pune in 2004 and his Ph.D. from SRTMU Nanded in 2012. He is working as Professor in Computer Engineering Department and Principal at Sinhgad Institute of Technology and Science, Maharashtra, Pune, India. He is having 24 years of experience in academics. His area of interest is Soft computing, Internet of Things, Text Analytics and Information management. He has published over 85 papers in national and international journals. He is a Member of IEEE, life member of International Association of Engineers, Indian Society of Technical Education and Computer Society of India. He is also a fellow of Institution of Engineers, India.